A Unified Approach to Automatic Indexing and Information Retrieval

Allen Ginsberg, AT&T Bell Laboratories

THE STRATEGY OF THE WORLDVIEWS SYSTEM IS TO UNITE AS MANY ASPECTS OF INFORMATION ORGANIZATION AS POSSIBLE AROUND ONE SIMPLE, FAMILIAR, YET VERSATILE KNOWLEDGE REPRESENTATION FRAMEWORK: A LATTICE-STRUCTURED VERSION OF THE TRADITIONAL THESAURUS.

ACCORDING TO CONVENTIONAL wisdom, we are living at the dawn of an information age, in which the united power of computers and high-bandwidth communications will give us immediate access to the entire corpus of human knowledge. At the same time, however, people are deluged daily by hundreds of electronic mail messages and thousands of articles on bulletin boards. Thus, we are also witnessing the beginning of information overload.

To reap the benefits of the information age without being jaded by overload, we need a coherent strategy for organizing vast quantities of information and providing easy-to-use retrieval tools. WorldViews is an experimental system built on such a strategy: to unite as many aspects of information organization as possible around one simple, familiar, yet versatile knowledge representation framework. It uses a lattice-structured version of the traditional *thesaurus* framework.[1,2] The term "thesaurus" refers here to a structured set of subject headings or descriptors, rather than a literary thesaurus, which is a list of words and their synonyms. Drawing on ideas and techniques from information retrieval and AI, WorldViews focuses on the unified use of a thesaurus to automatically index and retrieve information. Another aspect of information organization, the task of alerting users to relevant documents based on user interest profiles, can easily be encompassed by this strategy but is not discussed here.

Thesauri and related conceptual frameworks have long been a staple of library science and are used in numerous on-line databases.[2] However, these databases rely on human experts to index documents using terms from an associated controlled vocabulary or thesaurus. Moreover, few scientists use the logical structure of thesauri for retrieval or as a natural framework for exploring collections. While some research has concerned automatic thesaurus construction,[3] the idea of using on-line thesauri as a fundamental organizing principle in automatic indexing and information retrieval has received little attention. (The sidebar on pages 50-51 discusses related work in the AI community.)

Metaphorically, a thesaurus provides a conceptual map of the information space corresponding to its areas of application. WorldViews realizes this metaphor by integrating query-initiated search with subsequent navigation through a well-defined space of relevant subtopics. This space is not limited to statically related subtopics, but can include the dynamic calculation of new subspaces defined by previously unrelated conjunctions of thesaurus entries (described in more detail later).

WorldViews has been used to process electronic news articles as well as abstracts of technical reports from Bell Labs and other organizations. These documents

0885/9000/93/1000-0046 $3.00 © 1993 IEEE

cover a wide range of topics, ranging from relatively academic pieces about neural networks, to pragmatic reports about software requirements, to articles on affirmative action. Due to the proprietary nature of Bell Labs reports, the examples in this article use only non-Bell abstracts.

Overview of WorldViews

WorldViews consists of a system for automatic document indexing, an information retrieval system, and a user interface. These subsystems are unified in the sense that they all rely heavily on one on-line thesaurus. WorldViews is written in C, and uses X Windows for its user interface.

Automatic indexing is the process by which thesaurus entries are automatically assigned to documents as content descriptors. The automatic-indexing subsystem currently runs in batch mode as a separate process. Its input is a set of documents, *D*, and the thesaurus, which provides it with descriptors. The output is an updated version of the thesaurus: Any thesaurus entry that has been used to index a document in *D* will now include a pointer to, or a posting for, that document. Using a constrained form of spreading activation in a semantic network,[4] this subsystem traces connections among concepts to draw out a document's implicit conceptual content based on the explicit concept references it contains. For example, a document that explicitly uses the concepts "wasps" and "mosquitoes" but never uses the term "insects" will still be indexed under both the more specific and the more general terms.

Automatic indexing also estimates the percentage of a document's content that is relevant to each of the assigned descriptors. This is called the *content number* for that document with respect to a thesaurus element. Thus, in the previous example, the document's content number relative to the term "insects" will be computed by combining the content numbers associated with the narrower terms "wasps" and "mosquitoes."

When the same term represents more than one word sense, the indexing subsystem computes *distances* among thesaurus entries to choose the correct term to apply. For example, the word "programs" is a constituent of several entries in the thesaurus, including "computer programs," "social programs," and so on. Before using this term to index a document, the system will try to determine the correct sense of the term.

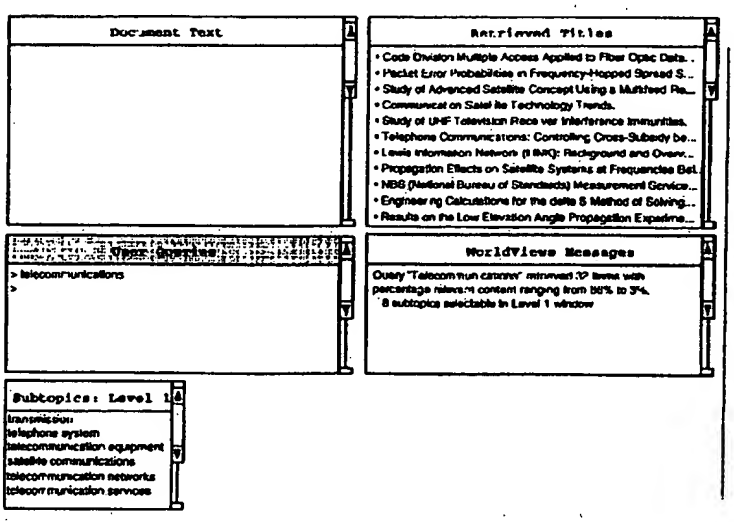

Figure 1. WorldViews display after the user types "telecommunications."

The indexing subsystem is integrated with portions of a Bell Labs information retrieval system called Slimmer.[5] Essentially, WorldViews uses Slimmer to create inverted files of terms that occur in document collections but are not contained in the thesaurus. This allows WorldViews to respond to user queries that it cannot fully interpret relative to its thesaurus. This simple expedient guarantees that, in the worst case, WorldViews' levels of recall and precision will be on a par with those of a traditional keyword system. Since most real-world document collections are continually expanding, while thesauri are updated only periodically, some approach of this sort is necessary to make WorldViews reliable in practical applications.

The information retrieval subsystem uses the thesaurus in a number of ways. First, it tries to interpret user queries as conjunctions of thesaurus entries. It also uses the postings associated with thesaurus entries to retrieve corresponding documents. When processing an interpreted query, the system uses the thesaurus to find related subtopics that will be incorporated as part of its response. As later examples will show, this capability helps brings the information space metaphor to life.

In the example in Figure 1, the user has typed in the query "telecommunications," which is echoed in the User Queries window. The system has responded in two ways. In the Retrieved Titles window, the system has listed the titles of documents that it has judged to be relevant (in descending order of the percentage of the document considered relevant). It has also printed a message in the Messages window saying how many titles are in the scrollable list, and the range of the percentage of relevant content. At the same time, in the window labeled Subtopics: Level 1, the system has displayed a scrollable list of more specific subtopics related to the user's query. Each item in this list applies to one or more documents in the collection, so these are all possible options.

The user can now choose to view a document by mouse-clicking on the title, or click on an item in the subtopic window. In our example, the user clicks on the subtopic item "transmission," which causes a new subtopic level (Level 2) to be displayed. The user now selects "signal interference" in this window, and then selects the title "Study of UHF Television Receiver Interference Immunities," which leads to the display shown in Figure 2. The messages in the WorldViews Messages

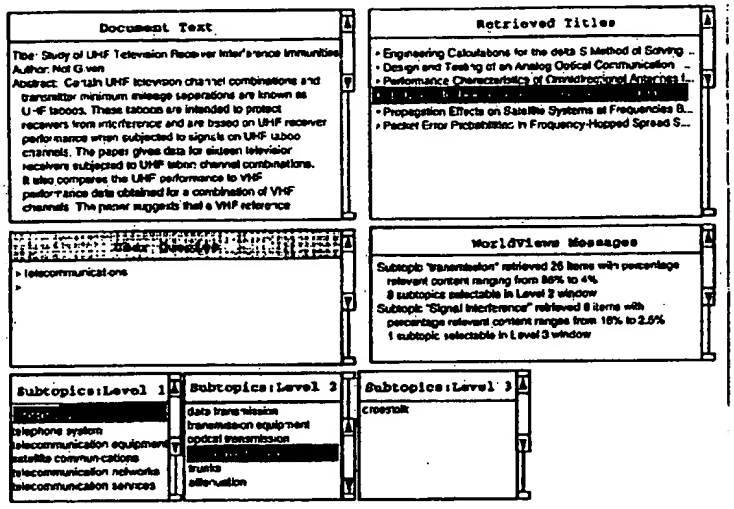

Figure 2. WorldViews display after the user makes subtopic and title selections.

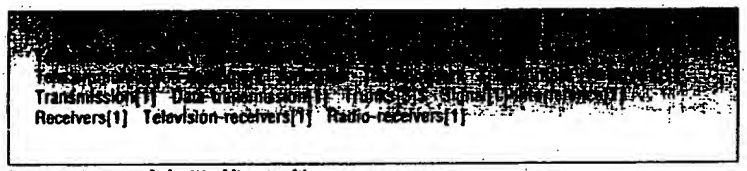

Figure 3. Format of the WorldLattice file.

window reflect the results of the user's two subtopic selections. (I'll use this document as an example throughout the article.)

The user can back up and select a different subtopic by clicking on any item in any subtopic window. In Figure 2, if the user were to select the Level 1 item "telecommunication equipment," the Level 3 subtopic window would disappear, and the Level 2 subtopic window would be repopulated with new subtopics. The row of subtopic windows shifts to the left (off the screen) if more space is needed to accommodate deeper levels of subtopics.

Lattice-structured thesauri

When I began work on the WorldViews project, I needed to obtain or build a lattice-structured thesaurus. Since the original documents were to be messages drawn from electronic bulletin boards, professionally constructed academic thesauri would likely be inadequate. For example, word sense information would be vital to disambiguation, but professionally constructed thesauri do not contain this information in machine-usable format, if at all. Thus the WorldLattice was created—my attempt to build a thesaurus for everyday life, so to speak. It currently consists of roughly 3,000 *nodes* (preferred terms), with an additional 500 or so equivalent terms or aliases. It also contains 391 lexical entries that are associated with two or more senses (nodes).

As the project's focus shifted to Bell Labs technical reports, the WorldLattice grew to resemble a technical thesaurus. Eventually, therefore, I obtained IEEE's Inspec thesaurus.[6] The latest edition of Inspec deals with the fields of physics, electronics, and computing. According to an Inspec representative, from 1973 through 1991 the thesaurus grew from 4,000 to 6,000 preferred terms (used in indexing) and from 3,000 to 7,000 lead-in terms (cross-references). This supports the view that thesauri are never complete, so that the supplemental use of an inverted file in a system like WorldViews will always be needed.

I converted the Inspec on-line thesaurus into WorldViews format and added some entries to let the system apply its word sense disambiguation technique. The modified thesaurus contains 620 lexical entries with two or more associated senses. (From here on, "Inspec" refers exclusively to the modified version.) All the WorldViews techniques have been successfully applied to the Bell Labs document collection using the Inspec thesaurus.

While this article focuses mainly on results obtained using the WorldLattice, WorldViews can be used with any lattice-structured thesaurus. (I will also present some results obtained using Inspec.)

Properties of lattice-structured thesauri. A lattice-structured thesaurus is a lattice of word or phrase *senses* or *concepts*. Mathematically speaking, a lattice is a partially ordered set in which every pair of elements has both a greatest lower bound and a least upper bound. Greatest lower bounds are of no interest in this context, but the idea of looking at the least upper bounds of two concepts in a lattice-structured thesaurus turns out to be useful, as we shall see later. (In contrast to strict mathematical usage, nonunique least upper bounds are allowed.)

Following thesaurus terminology, the nodes of a lattice-structured thesaurus are related via the *broader term* (BT) relation and its inverse, the *narrower term* (NT) relation.[2] Thus, for example, the concept of *physics* is a narrower term (more specific) relative to *the sciences*, while the latter is a broader term (more general) relative to the former. In this article, these relations are closed under the transitivity property. For example, if Z is narrower than Y, and Y is narrower than X, then Z is narrower than X. Viewed in this way, these relations are basically equivalent to the *more-general-than* and *more-specific-than* relations, more commonly used in AI work on subsumption.

In many thesauri, entries can be associated via the *related term* (RT) relation, which is distinct from BT/NT. While the BT/NT relationship is reserved for genus/species relationships, the RT relationship represents other forms of relatedness.[2] For example, we should use the RT relation to connect the terms "Vietnam War" and "Vietnam War veterans." However, it also makes

sense to regard the latter term as an NT with respect to the former, thus avoiding the use of RT.

Unlike the BT/NT relations, RT is *symmetric*; that is, if A bears the RT relation to B, then B bears it to A. This means that the RT relation does not induce a partial ordering over the elements in its domain and, therefore, does not generate a lattice. For this reason, and because many RT instances could be reformulated as BT/NT relations, earlier versions of WorldViews did not use RT at all (the current version does allow RT).

Thesauri generally offer some way to encode the fact that different character strings can refer to the same topic; for example, "AT&T," "American Telephone and Telegraph," and "American Telephone & Telegraph" all refer to the same entity. Thus, each WorldViews thesaurus entry has an official label and, possibly, aliases. (Aliases play the role of equivalent terms in thesauri.) When the indexing system starts, it compiles the labels and their aliases into finite state machines for text processing. WorldLattice node names, their aliases, and their constituent words are called *lexical entries*.

Figure 3 shows an example of WorldViews' original, simple mode of thesaurus definition, in this case a file that WorldViews read to create the WorldLattice (the current mode uses a database of records with field specifiers). Each line describes a set of links in the WorldLattice. The first node name in a line is the more general node, and the rest are more specific nodes. Some node names contain one or more bracketed numerals that differentiate which *sense* is meant: "transmission[1]" refers to telecommunications transmissions, while "transmission[2]" refers to vehicle transmissions. When WorldViews reads in the WorldLattice, it automatically generates lexical entries for the node names in this file. Any word with bracketed numerals will have the requisite number of senses associated with it.

In Figure 3, the first occurrence of the word "transmission" has a "S" after it, which specifies a *default sense*. As the name implies, the system assigns the default sense to a word if the word sense disambiguation procedure fails to pick a sense. Default senses are a powerful device, but must be used with caution. A default sense might need to be removed or changed depending on the documents being analyzed.

When WorldViews reads in the WorldLattice file, it creates a data structure for each node specified. Each data structure contains pointers to its parents (nodes whose topics are more general) and its children (nodes whose topics are more specific). The system also computes path information concerning the node's location in the WorldLattice, and stores this information in the data structure for the node. This allows WorldViews to perform fast WorldLattice searches and traversals and quickly determine subsumption relations and least upper bounds among nodes. Basically, the logical relationship between any two nodes can be determined without examining any other node in the WorldLattice.

Automatic indexing

WorldViews' automatic-indexing procedure has two phases: Generating a list of WorldLattice concepts explicitly referenced by the document, and generating a map of the document that also includes implicitly referenced concepts.

Phase one. WorldViews scans each document once, determines *explicit concept references*, and tries to resolve the intended senses of polysemous words (words with more than one meaning). A word or phrase *W* in a piece of text *explicitly references* a concept (node) *C* in the WorldLattice if *W* is a syntactic variant of *C*'s label or an alias for *C*, and if the author's intended use of *W* is identical to the intended sense of *C* in the WorldLattice. So, for example, the sentence "The cat is on the mat," contains an explicit reference to the WorldLattice node *cats*, but it does not contain an explicit reference to the more general node *mammals*. The text does, however, contain an *implicit* reference to this more general node. If a piece of text *W* contains an explicit reference to a WorldLattice concept *C*, it also contains implicit references to any node that is more general than *C*.

When scanning a document, WorldViews looks at each word in order and does a table lookup to see whether there is a WorldLattice label or alias that the word matches or for which it could be the beginning. For example, the word "American" could be a direct match into a lexical entry, or it could be the beginning of a phrase that matches an entry such as "American Telephone and Telegraph" or "American Express." When the system does a table lookup on "American," it finds that a representation of a finite state recognizer is associated with this word. The system will use this recognizer to continue scanning until either some phrase is matched (the system looks for the longest possible matches), or no further match is found. Thus, if the system were scanning the phrase "American's in foreign ...," the system would match the word "American's" to "American," and the finite state machine would terminate once the word "in" is scanned, since no WorldLattice node is labeled with a phrase beginning "American in." As mentioned in the earlier overview, any word that fails to be matched and is not on the stop list will be inserted in an inverted index.

Thus, the lexical and syntactic analysis of the WorldViews text processor is fairly simple. Each call to the scanning procedure returns a structure containing pointers to the next lexical entry recognized in the text together with a pointer to the associated WorldLattice entry if the entry is not polysemous. If the entry is polysemous, the structure contains pointers to all the possible WorldLattice senses.

Using locality for word sense disambiguation. WorldViews uses the information gathered during the first phase to accumulate evidence concerning the intended sense of polysemous words. Suppose WorldViews is processing a document and finds a word or phrase that it knows has multiple meanings. For example, "interference" in the document in Figure 2 has two senses in the WorldLattice: One is the sense used in "electromagnetic interference," and the other is used in "there was interference on the play" with respect to football. However, the system must really choose among *six* possibilities to resolve the polysemy. The four other possibilities arise because "interference" is a constituent in other WorldLattice nodes: signal interference, wave interference, and electromagnetic interference use the first sense of the word, and pass interference uses the second. WorldLattice nodes whose labels share a common constituent sense are called *subsenses* of that sense.

In the current implementation, the system can resolve an ambiguity only in terms

Related work

Since the body of related work in automatic indexing, word sense disambiguation, and information retrieval is so large, this discussion focuses mainly on recent research in the AI community.

Natural language processing. Much of the AI work in all three areas relies to some extent on the use of natural language processing techniques, in particular, parsing algorithms. The Scisor[1] and Ferret[2] systems represent this school of research. Using frame-based knowledge representation techniques, they try to generate rich semantic representations of text. This makes it easier to develop natural-language question-answer user interfaces, which the authors claim yield higher rates of precision and recall. However, such systems typically deal with specialized domains. Scisor, for example, analyzes financial news stories concerning mergers and acquisitions; the system has been used in other domains, but still rather narrow ones.[1]

Aside from simple finite-state recognizers, WorldViews uses no parsing or related natural language processing techniques. In principle, such techniques could be integrated into WorldViews, either as a front end to the information retrieval system, or into the automatic-indexing system. However, since WorldViews' scope is essentially unrestricted, the effort involved in gathering and maintaining the necessary domain knowledge to make natural language techniques worthwhile would be prohibitive.

Spreading activation in semantic nets. Another major theme in AI work in these areas is the use of spreading activation in semantic networks. The Grant system, for example, uses this technique to match descriptions of relevant funding agencies to descriptions of grant proposals.[3] This is similar to WorldViews' use of its thesaurus to derive a document's implicit conceptual content from what the system has determined to be its explicit conceptual content. Grant does not do automatic indexing or word sense disambiguation.

A key difference between Grant and WorldViews is in the nature of the constraints applied to the spreading activation technique. Grant uses distance and fan-out restrictions as well as rule-like constraints, called *path endorsements*, to limit the nodes that can be activated. WorldViews' constraint on spreading activation is based entirely on the logical structure of the thesaurus and the nature of the task at hand. Thus, during automatic indexing, when the system uses spreading activation to establish a document's implicit conceptual content given its explicit content, the activation can only go upward from more specific to more general terms.

Hirst and Charniak's use of marker passing in word sense disambiguation[4] is related to WorldViews' use of locality among thesaurus entries to determine word or phrase senses. A key difference is that the WorldViews technique relies exclusively on the distances between entries and their interconnectedness in the thesaurus. In the case of Hirst and Charniak, each node in the semantic net is a frame containing linguistic and conceptual knowledge to be used in the disambiguation process. Also, WorldViews bases its decisions about word sense on a wider context, essentially paragraph-sized, while Hirst and Charniak use a sentence-sized context.

Belew's connectionist approach to information retrieval[5] is another example of work that is related to the WorldViews approach.

Machine-readable dictionaries in word sense disambiguation. A number of approaches to word sense disambiguation use machine-readable dictionaries.[6] Slator advocates processing dictionary entries to extract a semantic network, which in turn is used in the disambiguation task as part of the overall natural language processing task.[7] Other researchers have developed approaches that work directly from a dictionary's contents.[8,9] Lesk, for example, counts the number of words shared by dictionary definitions of the competing senses of polysemous words.[8] This overlap is taken as evidence in favor of those senses.

On the surface, the approaches of Lesk and of Wilks and colleagues[9] seem very different from WorldViews'. Nevertheless, the intuitions underlying these approaches appear to be similar. According to Wilks and colleagues, word senses that occur together locally in a document are likely to be semantically related. For both sets of researchers, the degree of semantic relationship is measured by word co-occurrence statistics relative to dictionary definitions; WorldViews measures the degree of

of an actual WorldLattice node. The two general senses of "interference" do not occur as independent nodes, but only as constituents of the four nodes just mentioned (that is, as subsenses). Thus, only four choices exist after all. However, there are situations in which the evidence supports selection of a sense, but not of a subsense. For example, it is often easier to determine that a text is using "interference" in the sense of "signal interference," "wave interference," and "electromagnetic interference," than it is to determine which of these specific subsenses is intended. Indeed, some other subsense, not included in the WorldLattice at all, might be intended. Allowing the system to choose a general sense in such situations will lead to improved performance. This possibility will be incorporated in future versions of the system.

WorldViews uses two techniques to resolve polysemy. The first is quite simple: If the system sees a polysemous word in a known nonpolysemous phrase, then it resolves the ambiguity immediately (for that occurrence of the word). For example, "UHF" is an alias for the WorldLattice node *ultra-high frequency*, and "VHF" is an alias for the node *very high frequency*. The word "frequency" is ambiguous, but the phrases in question are not, so the system can resolve these potential ambiguities immediately.

If WorldViews cannot resolve an ambiguity in this way, it waits until the document has been completely scanned. While scanning, it keeps track of all unresolved ambiguities, gathering evidence that might contribute to their eventual resolution. WorldViews' policy of deferring decisions is meant to be used only with short documents such as abstracts or electronic news messages. Full-length documents such as papers and manuals are broken into groups of paragraphs totalling at least 150 words, and then indexed separately.

The evidence that WorldViews accumulates concerning an ambiguous word or phrase W is the WorldLattice distances between the nodes for the various senses/

semantic relationship by distances between senses relative to the thesaurus.

In contrast to these knowledge-based approaches, recent work by Church and colleagues uses purely empirical statistical analysis of text to construct (and train) word sense discriminators.[10]

Rule-based and expert system approaches. The Rubric system uses a rule-based approach to information retrieval.[11] A collection of multilevel weighted rules is used to define a topic or concept. Rules at the lowest level define patterns of strings that are used to search text; higher rule levels relate these patterns to intermediate concepts or abstractions needed to define the topic. Rule weights are used to compute relevance rankings. Rubric is not used for automatic indexing, nor does it explicitly deal with word sense disambiguation.

Rubric's rules are intended to capture experiential or operational knowledge of the sort that is contained in a successful, complex Boolean query or search strategy. Knowledge at this level tends to consist of varied bits and pieces, rolled together into a complex structure. For example, the World Series concept[11] is defined in terms of both world knowledge rules (for example, *Saint Louis Cardinals is a baseball team*) as well as linguistic rules (such as *"St." can abbreviate "Saint"*). Thus, even though the Rubric rule set grows larger and larger, it really never contains a purely declarative knowledge representation corresponding to WorldViews' thesaurus. Because it uses this kind of knowledge representation, WorldViews can support features such as subtopic breakdowns, conjunctive-expression expansion, and the automatic integration of the results of such expansions into its thesaurus structure.

Croft and Thompson have looked at using expert systems as intelligent assistants or intermediaries for information retrieval systems.[12] Gauch and Smith also consider the role of a thesaurus in automatic query reformulation by an expert system intermediary.[13]

References

1. P.S. Jacobs and L.F. Rau, "Scisor: Extracting Information from On-Line News," *Comm. ACM*, Vol. 33, 1990, pp. 88-100.
2. J. Carbonell, M. Mauldin, and R. Thomason, "Beyond the Keyword Barrier: Knowledge-Based Information Retrieval," *Information Services and Use*, Vol. 7, No. 4-5, Mar. 1987, pp. 103-117.
3. P.R. Cohen and R. Kjeldsen, "Information Retrieval by Constrained Spreading Activation in Semantic Networks," *Information Processing and Management*, Vol. 23, No. 4, 1987, pp. 255-268.
4. G. Hirst and E. Charniak, "Word Sense and Case Slot Disambiguation," *Proc. Nat'l Conf. Artificial Intelligence*, Morgan Kaufmann, San Mateo, Calif., 1982, pp. 95-98.
5. R.K. Belew, "A Connectionist Approach to Conceptual Information Retrieval," *Proc. First Int'l Conf. Artificial Intelligence and Law*, ACM, New York, 1987, pp. 116-126.
6. R. Krovetz and W.B. Croft, "Word Sense Disambiguation Using Machine-Readable Dictionaries," *12th Annual Int'l Conf. Research and Development in Information Retrieval*, ACM, New York, 1989, pp. 127-136.
7. B.M. Slator, "Sense and Preference," *Computer and Mathematics with Applications*, Vol. 23, No. 6-9, Mar.-May 1992, pp. 391-402.
8. M. Lesk, "Automatic Sense Disambiguation Using Machine-Readable Dictionaries: How To Tell a Pine Cone from an Ice Cream Cone," *Proc. SigDoc*, ACM, New York, 1986, pp. 24-26.
9. Y. Wilks et al., "Providing Machine-Tractable Dictionary Tools," *Machine Translation*, Vol. 5, No. 2, June 1990, pp. 99-154.
10. K.W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, Vol. 16, No. 1, Mar. 1990, pp. 22-29.
11. R.M. Tong et al., "Conceptual Information Retrieval Using Rubric," *Proc. 10th Annual Int'l Conf. Research and Development in Information Retrieval*, ACM, New York, 1987, pp. 247-253.
12. W.B. Croft and R.H. Thompson, "I³R: A New Approach to Natural Language Processing for Document Retrieval Systems," *J. Amer. Soc. for Information Science*, Vol. 38, No. 6, Nov. 1987, pp. 389-404.
13. S. Gauch and J.B. Smith, "Search Improvement Via Automatic Query Reformulation," *ACM Trans. Information Systems*, Vol. 9, 1989, pp. 249-280.

subsenses of W and the nodes for the sense/subsense of unambiguous words (or senses/subsenses of other ambiguous words) in the document. These distances are translated into scores, with shorter distances yielding higher scores. After all the evidence is in, the sense/subsense with the highest score is selected if

(1) the score surpasses a certain threshold (currently set at 3), and
(2) the difference between this score and the score for the next highest competing sense/subsense is greater than the latter's score.

Thus, if the highest score is 6 and the next highest score is 4, no decision will be made unless the word has a default sense/subsense, in which case the default will be chosen.

The intuition behind this approach is that the WorldLattice nodes representing the word senses in a document are often huddled together rather than spread apart. Consider the sentence "The program has courses that are interdisciplinary in nature." Intuitively, the sense of the word "program" corresponding to "computer program" has nothing to do with the sense of the word "courses" corresponding to "meal." This is reflected in the large distance separating the nodes corresponding to these senses in the WorldLattice. On the other hand, the sense of the word "program" that corresponds to "academic program" and the sense of the word "courses" that corresponds to "curriculum" do have something to do with one another. In the WorldLattice, these concepts are siblings under the concept *curriculum*.

So, we can interpret ambiguous words by looking for word senses whose nodes minimize the sum of the distances between the selected nodes (including the nodes for the document's unambiguous words). However, the actual technique used is simpler and more efficient. It is based on the notion of a *least upper bound* of two WorldLattice concepts. Let c, c_1, c_2 be three WorldLattice nodes. There is a *downward path* from c_1 to c_2 if $c_1 = c_2$, or if c_2 can be reached from c_1 by following a sequence of

Table 1. Disambiguation of the word "interference." The three competing subsenses for this abstract are subsense1, signal interference (communications); subsense 2, wave interference (physics); and subsense 3, pass interference (football).

Compared with	Ambiguous?	Least upper bounds	Max distance	Change in score
UHF	No	Transmission	2	+3 to subsense 1
		Wave phenomena	2	+3 to subsense 2
Television receivers	No	Telecommunications	2	+3 to subsense 1
Television channels	No	Telecommunications	3	+3 to subsense 1
Receivers	Yes	Football	2	+1 to subsense 3
		Telecommunications	2	+1 to subsense 1
Transmitters	No	Transmission	2	+3 to subsense 1
Channels	Yes	Transmission	1	+2 to subsense 1
Signal	Yes	Transmission	1	+2 to subsense 1
VHF	No	Transmission	2	+3 to subsense 1
		Wave phenomena	2	+3 to subsense 2
Total score for subsense 1 = 18		Total score for subsense 2 = 6		Total score for subsense 3 = 1

more-specific-than links. Now, node c is a least upper bound of c_1 and c_2 if and only if

(1) there is a downward path p_1 from c to c_1,
(2) there is a downward path p_2 from c to c_2, and
(3) any other node c' that satisfies (1) and (2) is not in the paths p_1 or p_2.

The paths p_1 and p_2 are called *legs*; the number of links traversed along a leg is the distance associated with that leg.

Table 1 shows the scoring process for "interference" in our example. The ambiguous word is compared once against every word in the document. In particular, comparing "interference" with "UHF" involves computing the distance from each of the three nodes corresponding to the three subsenses of "interference" to the single node *UHF*. As a measure of the distance, we take the maximum one-leg distance to the nearest least upper bound. The largest distance for which an increment will be generated is 3. Scores for comparisons against known unambiguous words are higher than for comparisons against ambiguous words. A distance of 1 when compared with an unambiguous word yields an increment of +5, whereas the same distance with respect to an ambiguous word yields an increment of +; a distance of 2 yields increments of +3 and +1, respectively, and a distance of 3 yields increments of +1 and +1, respectively. A distance of 0 can occur if an unambiguous phrase using an ambiguous word (for example, "television receivers") appears with an ambiguous occurrence of that word (such as "receivers") in the document. As shown in Table 1, the distance from the node *signal interference* to the nearest least upper bound with *UHF*, via *transmission*, has a distance of 2. Since the term "UHF" is known to be unambiguous, this comparison yields an increment of +3 to the score for the corresponding subsense of "interference" (subsense 1). The second subsense, "wave interference," also has a least upper bound with *UHF*, namely *wave phenomena*, with a maximum one-leg distance of 2. None of the scores for the other subsenses of the word are incremented as a result of this comparison because the distances are too large. In this case, an increment of +10 is awarded to the sense/subsense in question.

Let's look at one final detail of this procedure. The results of comparisons with more specific subsenses of a word or phrase are considered only when they yield distinct least upper bounds from those derived from comparisons with any more general sense/subsense. Thus, in our example, the term "electromagnetic interference" is a child of "wave interference," but every comparison with either term yields exactly the same least upper bound. Intuitively, this means there is no independent evidence supporting the more specific subsense as opposed to the more general one.

The technique I have described involves choosing values for a number of parameters: for example, the size of the scores, the maximum distances allowed, and so on. The absolute values assigned to the parameters appear far less important than their relative sizes. For example, smaller distances between senses or subsenses should yield higher scores, and comparisons with senses or subsenses known to occur in the text should yield higher scores than comparisons with polysemous senses or subsenses.

Experimental results. I evaluated this word sense disambiguation technique using about 12,000 Bell Labs abstracts. I gathered two types of statistics: general performance, and performance on specific examples of ambiguous words. I also collected experimental evidence concerning the change in performance caused by changes to thesauri that were made to correct specific performance problems. The system detected 39,138 polysemous word or phrase occurrences, and picked a word sense in 32,157 of these cases, or about 82 percent of the time. (This number drops to 72 percent if default senses are not used). Only 1,914 of these cases were decided immediately on the basis of a polysemous word being part of a larger single-sense phrase; the rest required the use of WorldLattice locality, as described earlier.

How often did the system make a good choice? Obviously, it was not practical to check all 32,157 cases. Instead, I selected and checked a random sample of 200 cases. The system made good word sense choices 77 percent of the time over this sample, which can be extrapolated to the entire set with an error range of ± 5 percent. This general level of performance is comparable to those reported elsewhere.[7,8]

However, the current WorldLattice is incomplete relative to the set of documents indexed. We would really like to know the level of performance that can be expected by a version of WorldViews that uses a relatively complete WorldLattice with respect to the documents indexed. While I cannot answer this question at this time, the following experiment shows that the system's performance tends to improve as the WorldLattice is built up. After analyzing 50 electronic news messages, the system detected 153 polysemous word occurrences and resolved 102. I checked a random sample of 20 of these decisions: 13 were good choices, and seven were poor. So far, the experiment has the same form as the one involving the much larger collection of abstracts. Using this smaller sample, however, we can experimentally address the issue of convergence. Suppose we try to correct the seven poor decisions by expanding or altering the WorldLattice in appropriate ways; what effect would these changes have on the system's performance on other

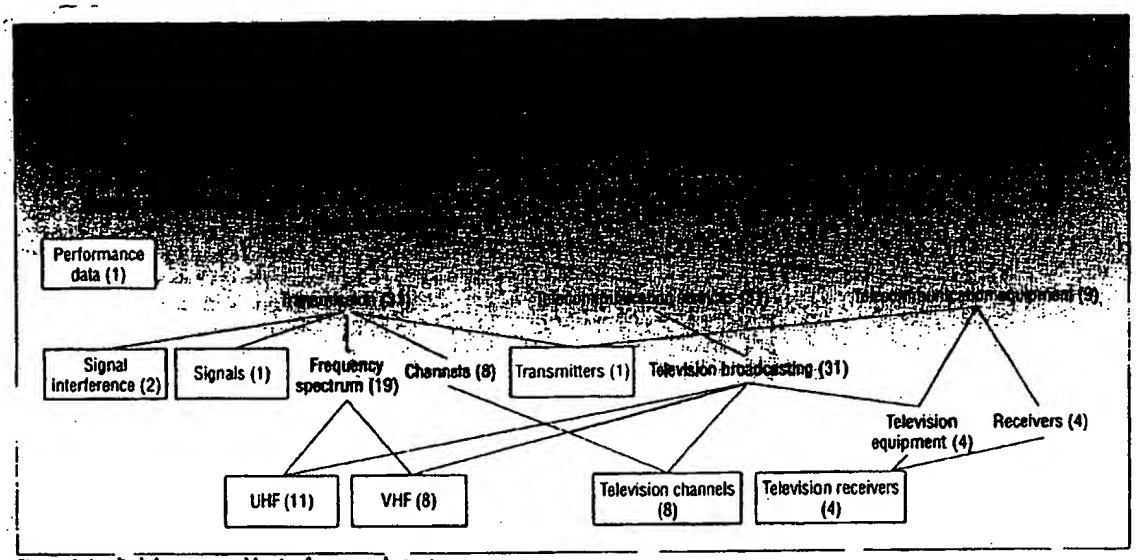

Figure 4. Implied document sublattice for example text.

cases? First of all, the seven problem cases were easily corrected by expanding the WorldLattice contents. I then ran the system over the same 50 messages using the updated WorldLattice, and compared the differences in log files from the two runs: Not only were the seven original bad decisions corrected, but 14 additional bad decisions were corrected as well. In no case did the system go from making a good word sense decision to making a bad one.

Thus, expanding the WorldLattice to correct problems will generally not cause problems with cases that the system was already handling correctly. Therefore, the level of performance achieved with the current WorldLattice is only a lower bound on the level of performance that can be achieved using this technique. As the WorldLattice comes to represent a more complete and accurate mapping of the domains of the documents it analyzes, the level of its word sense disambiguation performance should increase correspondingly.

This expectation has also been borne out by experiments with particular ambiguous words. For example, the word "current" has two senses in the Inspec thesaurus. Initially the system performed at about a 70-percent level of accuracy for occurrences of this word in the Bell Labs collection of 12,000 abstracts. Adding a number of new entries to Inspec to fix these problems brought the level of correct word sense decisions up to 90 percent for this word.

Phase two. The system uses the explicit concept references found in the first phase to generate a map of the document that also includes the concepts only implicitly referenced by the text. Since this map will be a proper part of the WorldLattice, it is called the *implied document sublattice*.

The implied sublattice is computed by beginning with the explicit concept references and working up, from NT to BT nodes. Each node tracks the number of times it has been visited, and updates its frequency with every visitation. (The system uses these values to estimate the relevant content of retrieved documents in response to user queries.) Since every upward chain of nodes in the WorldLattice terminates at the root node, and there are no cycles in this thesaurus, the process must halt. The frequency computed for the root node is called the *total content*.

Once this upward activation is complete, the system indexes the document by traversing the sublattice downward: Excluding the root node, the system starts by finding the set of most general implied nodes in the sublattice that were visited at least two times on the way up. Each of these nodes and all their descendants are used to index the document. As each of these nodes is visited, its percentage of relevant content is computed by dividing its frequency by the total content and multiplying by 100. These numbers, the *content numbers* for the corresponding nodes, are stored on disk with the document postings for use by the information retrieval subsystem.

Figure 4 shows the implied sublattice for our example document. The nodes in boxes are the document's explicit concept references; the number in parentheses is the concept's frequency of occurrence in the text. Nodes not in boxes are implied concept references; their frequencies are simply the sum of the numbers for the nodes below them.

Information retrieval

As the earlier examples showed, WorldViews' approach to information retrieval is based on using the WorldLattice as a conceptual map of an information space. The key task of WorldViews' information retrieval system is to interpret a user's query relative to the WorldLattice, and then access and (if needed) compute the subspace relevant to the query. The system also facilitates iterative user navigation through neighboring subtopics.

Query interpretation. Queries are interpreted via a simple inverted file search relative to the WorldLattice (or other indexing thesaurus), together with a subsumption analysis to produce the list of the most general thesaurus entries that match the given terms. Suppose, for example, the query is "elementary particles." The system will initially look for every thesaurus entry containing these terms. In

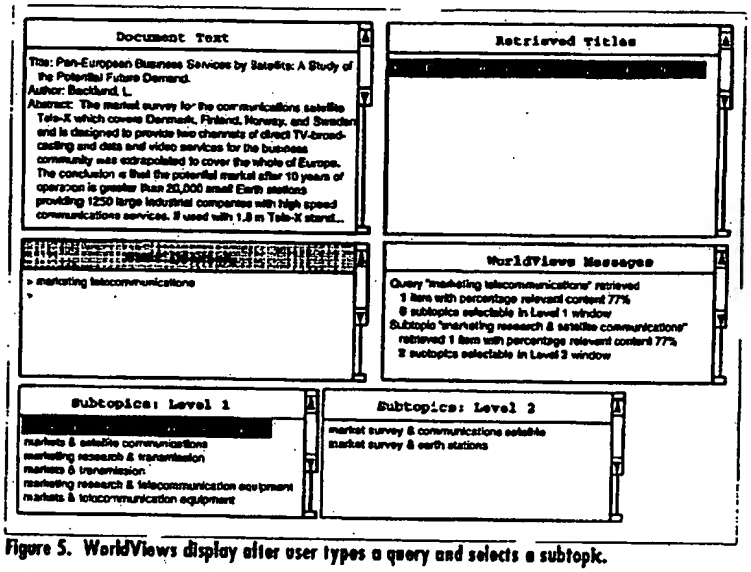

Figure 5. WorldViews display after user types a query and selects a subtopic.

the process, any entry that is determined to be subsumed by a more general entry found earlier will not be included. For example, Inspec contains the entries "elementary particle interactions" and "elementary particle weak interactions." The latter is narrower than the former, and so only the former entry will be included in the initial response to the query. The latter entry is, of course, accessible via subtopic expansion.

Using inverted file search allows a great degree of flexibility in query entry; word order, for example, has no effect on the response to a query. This also lets users type partial terms with identical results: typing "elem parti" will yield the same results as "elementary particles."

A more subtle, but equally or more important, advantage of using inverted file search is its ability to handle ambiguous queries. When a user enters an ambiguous query, the system simply returns all the most general matching entries for every possible sense. The returned topics contain all the information needed for the user to select the entry or entries that agree with the intended word senses. (Identical ambiguous entries are distinguished by displaying the names of their more general entries along with them.)

In forming its interpretation of an information request, the system constructs the most specific interpretation possible. For example, if a user types in the query "physics quantum mechanics," the system will ignore the broader term "physics." Thus the interpreted query consists solely of "quantum mechanics," and the system will return the thesaurus entry for that topic.

Conjunctive expressions. In the example in Figure 5, the user has typed the query "marketing telecommunications." WorldViews interprets this as a *conjunctive expression*, that is, a conjunction of two known thesaurus entries, "marketing" and "telecommunications." These terms are logically independent: neither subsumes the other in the thesaurus, that is, neither term is reachable from the other following the links in the thesaurus *in a single direction*. If the terms were not independent, only the more specific term would be used in processing the query.

The system handles conjunctive expressions in two ways. If it has already seen and processed the expression, those results are still available and are retrieved (I will elaborate on this shortly). For now, assume that the query is entirely new to the system. To generate a response, the system computes the intersection of the document postings contained under the terms in the interpreted query. The resulting list is displayed in the Retrieved Titles window (see Figure 5). When computing relevant content percentages for documents retrieved by conjunctive expressions, the information retrieval system uses the minimum of the content numbers associated with the document. In this example, the single item retrieved has a 7.7-percent content for "marketing" and a 34.6-percent content for "telecommunications," resulting in a 7.7-percent content with respect to the conjunctive expression.

Computing expansions of conjunctive expressions. To give the user more specific subtopics, the system needs to expand its thesaurus around the conjunction "marketing & telecommunications." Such expansions are calculated as needed, either in response to a query interpreted as a conjunction of logically independent WorldLattice nodes, or in response to a subtopic selection using the mouse. The six new conjunctive expressions generated by this process are listed in the Subtopics: Level 1 window in Figure 5; they represent all the *closest* logically minimal conjunctive expressions subsumed by "marketing & telecommunications" that actually describe documents in the collection. The notion of closeness of conjunctive expressions is defined as follows. Let E, A, and B be conjunctive expressions. A is closer to E than B if and only if E subsumes A and either E does not subsume B or A subsumes B. Using the example in Figure 5, if E equals "marketing & telecommunications," A equals "marketing research & satellite communications," and B equals "market survey & communications satellite," then the terms of the definition are satisfied.

The user now selects the subtopic "marketing research & satellite communications." This causes the system to iterate the preceding steps with respect to this particular conjunctive expression; that is, the document titles in the relevant intersection are computed and displayed, and the system expands its thesaurus around the expression to yield two new (more specific) expressions in the Subtopics: Level 2 window. The user then selects the single displayed title to bring up the text.

The expansion technique is fairly straightforward (see the schematic example in Figure 6). Suppose the user has issued a query that the system interprets as a new conjunctive expression "A & B," where A and B are logically independent thesaurus entries. For each term in the conjunctive expression, the system finds all the nearest, more specific thesaurus entries that contain document postings. For example, in Figure 6 the relevant terms for A are $A21$, $A12$, and C. $A22$ is not included because $A12$ is a closer node to A along the same path; $A13$ contains no document postings.

The system then goes through all the postings associated with these entries. For each document, it keeps track of all the entries that post it ("Applicable nearest terms" in Figure 6). Only documents 1 and 2 are posted by subterms falling under both *A* and *B*. This accounts for the new conjunctive expressions generated in the following line in the figure: two conjunctive expressions have been generated for document 2. Finally, the term *C*, which is a narrower term of both *A* and *B*, will be included in the expansion of "*A* & *B*" even though it is not a conjunctive expression itself.

As mentioned earlier, once WorldViews has expanded its thesaurus around a new conjunctive expression, the results are automatically integrated into its thesaurus structure, so that the system does not need to duplicate work. The system can do this because the logical relationship between any two conjunctive expressions, or between a conjunctive expression and a simple thesaurus entry, is completely determined by the known logical relationships among the thesaurus entries used in conjunctive expressions.

While this procedure is straightforward, it can be expensive to carry out: the complexity of the worst case is governed by the product of the largest number of independent subterms that can occur under each term to be expanded. Simply dealing with a large number of postings can slow down performance. Nevertheless, experience with 12,000 Bell Labs abstracts has been quite reasonable. Computing the expansions for relatively specific queries, for example, "quantum well & quantum tunneling & VLSI," leads to no slowdown in performance. Computing a general expansion such as "telecommunications & Asia," takes about 5 to 10 seconds.

While more experience with larger collections is needed to better judge the average compute times for expansions, it is unlikely that the technique described here can be applied uniformly in real time. Depending on the circumstances, however, numerous satisfactory options are available. For example, the user could be invited to select a particular conjunctive expression from a menu, rather than waiting for the system to compute all the conjunctive expressions generated by expansion of some expression. Alternatively, the system could carry out the computation during times of minimal usage and notify the user on completion.

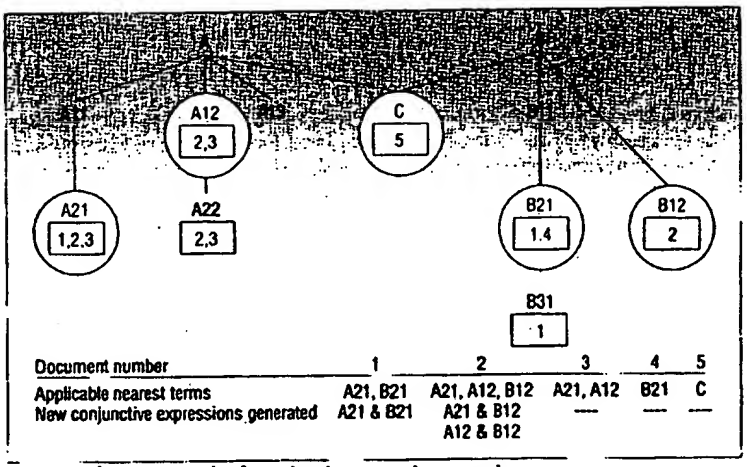

Figure 6. Schematic example of a conjunctive expression expansion.

The results of the experiments with WorldViews compare favorably with other techniques. They also indicate that better performance can be expected with a more mature thesaurus.

General versus specific queries. These examples show that WorldViews gives reasonable, well-structured responses to general queries such as "telecommunications." With large collections of Bell Labs documents, this and similarly general queries cause WorldViews to access thousands of document titles. As one would expect, addressing such queries to a keyword system such as Slimmer returns a very small fraction of the applicable documents. The reason, of course, is that most technical documents about a specific subtopic in telecommunications never explicitly use the word "telecommunications."

What makes the WorldViews response to such queries reasonable, however, is not the large number of relevant titles returned, but the inherently structured nature of the response. The user does not need to scrutinize numerous titles or documents; the subtopics navigation mechanism allows the user to easily descend to more specific levels, or back up to follow other paths. The list of retrieved titles changes accordingly; as more and more specific subtopics are selected, fewer and fewer titles remain. Documents with low relevance to the initial general query, but with higher relevance to the more specific selected subtopics, will move up the list of remaining titles, as can be seen by comparing the titles in Figures 1 and 2.

What about the opposite problem: Can WorldViews generate a reasonable response to *specific* queries? Typically, the responses to such queries contain few, if any, document titles. Nevertheless, the collection might include closely related documents that would also address the user's underlying information needs. For example, returning to Figures 1 and 2, if the user had begun by issuing the relatively specific query "crosstalk," then only one document title would be retrieved, with no subtopics. However, "crosstalk" is a narrower topic falling under the broader topic of "signal interference." If "signal interference" or any narrower subtopic under it contains additional document postings, we should make those possibilities apparent to the user by including the additional items and subtopic windows in the display, and printing an appropriate message. (This will be included in future versions of the interface.)

Hybrid technique. If none of the query can be interpreted relative to WorldViews' thesaurus, the entire character string is handed to Slimmer's keyword search facility, and the results are returned to WorldViews for display. Of course, only titles, text, and messages can be displayed, since no thesaurus subtopics can be retrieved. Suppose, on the other hand, that part of the query is known to WorldViews and part is not. Say, for example, that "telecommunications" is known to WorldViews, but "marketing" is not. In this case, the system would hand the unknown part to Slimmer, which would use its inverted index and return a list of postings, that is, pointers to

documents containing the term. If more than one term is given to Slimmer, the intersection of the postings is returned. In the meantime, WorldViews handles the rest of the query as though the other portion had not been given at all. If Slimmer returns no postings, then the whole query fails. Otherwise, as before, WorldViews computes the intersection of the document postings it finds with those returned by Slimmer. Thus, in this example, WorldViews first computes the intersection of the postings for "marketing" with those for "telecommunications" and lists the results in the Titles window. It also computes this intersection for each Level 1 subtopic found for "telecommunications," for example, "marketing & satellite communications." Thus, in the case of a partially interpreted query, WorldViews can still perform partial expansions of conjunctive expressions, but the unknown portion of the query remains the same in all the subtopics generated in all levels.

Performance

The rate at which the automatic-indexing system processes text depends on the properties of the lattice-structured thesaurus being used to do the indexing. Clearly, a thesaurus with greater average depth will tend to increase the amount of processing needed to construct the implied document sublattice; average node connectivity is also a factor here. The number of nodes in the lattice (and the number of aliases) will also affect processing time, since the more nodes in the lattice, the more likely it is that concepts will be found in the text.

However, even in the worst case (where every concept in a document somehow forces the algorithm to consider every other node in the indexing lattice), the amount of processing time grows only linearly with the number of concepts. This is also true of the word sense disambiguation part of the algorithm.

Experimental results also confirm that the indexing algorithm is fast. When using the WorldLattice as the indexing thesaurus, the algorithm processes text at the rate of about 30 million characters an hour (in processing time) on a machine performing 25 million instructions per second. I derived this figure by processing roughly 12,000 Bell Labs abstracts containing about 15 million characters of text in about 30 minutes. In comparison, just creating an inverted index for the same collection takes about 15 minutes of processing on the same machine. Results using Inspec as the indexing thesaurus are slightly faster, roughly 45 million characters per hour. Even though Inspec contains more than twice as many nodes as the WorldLattice, it is on average a shallower lattice, and this accounts for the difference in processing time. The average distance of a node to the WorldLattice root is 8.230, while the average distance of a node to the root of the Inspec lattice is 2.659.

During indexing, the system keeps a representation of the entire thesaurus in memory. The space required is about 4 Mbytes for the WorldLattice and about 8 Mbytes for Inspec. We foresee no problems in ultimately accommodating thesauri of up to 20,000 entries, while maintaining a comparable rate of throughput. Since query interpretation proceeds via an inverted file approach, entire thesauri do not need to be memory-resident during retrieval, but are brought into memory as needed.

HISTORICALLY, WORK IN INformation retrieval has emphasized two measures of retrieval effectiveness: recall and precision. While the system outperforms traditional keyword-based retrieval systems, comparisons with state-of-the-art information retrieval approaches have yet to be undertaken. Plans are underway to compare it to the Smart system, which uses the vector space approach,[9] or the Inquery system, which is based on the inference net approach.[10]

With the increasing emphasis on the role of moving and managing information in today's world, information retrieval should focus not only on techniques for improving precision and recall, but also on techniques and systems that educate. In other words, information retrieval should provide not only better access to documents for users who have a good idea of what they are looking for, but also useful and knowledge-imparting responses to users who need to learn about some topic or area with which they have little or no familiarity. WorldViews is intended to further that mission.

Acknowledgments

I thank the reviewers for their critical comments on this work. I thank Les Lunas and Bob Waldstein for access to Linus databases and the Slimmer system, and their comments on and interest in this work. I also thank David Lewis for his comments and suggestions.

References

1. M. Bullard, H. Landau, and A. Tamir, "The Hierarchical Indented Thesaurus System," *Proc. ASIS Annual Meeting*, 1971, pp. 373-380.
2. F.W. Lancaster, "Thesaurus Construction and Use: A Condensed Course," tech. report, Unesco, 1985.
3. G. Salton, "Experiments in Automatic Thesaurus Construction for Information Retrieval," *Information Processing*, Vol. 71, 1972.
4. P.R. Cohen and R. Kjeldsen, "Information Retrieval by Constrained Spreading Activation in Semantic Networks," *Information Processing and Management*, Vol. 23, No. 4, 1987, pp. 255-268.
5. R.K. Waldstein, "Slimmer: A Unix-System-Based Information Retrieval System," *Reference Services Rev.*, Vol. 16, 1988, pp. 69-76.
6. *Inspec Thesaurus*, IEEE Service Center, Piscataway, N.J., 1991.
7. M. Lesk, "Automatic Sense Disambiguation Using Machine-Readable Dictionaries: How To Tell a Pine Cone from an Ice Cream Cone," *Proc. SigDoc*, ACM, New York, 1986, pp. 24-26.
8. Y. Wilks et al., "Providing Machine-Tractable Dictionary Tools," *Machine Translation*, Vol. 5, No. 2, June 1990, pp. 99-154.
9. G. Salton, ed., *The Smart Retrieval System*, Prentice-Hall, Englewood Cliffs, N.J., 1971.
10. H. Turtle and W.B. Croft, "Inference Networks for Document Retrieval," *13th Annual Int'l Conf. Research and Development in Information Retrieval*, ACM, New York, 1990, pp. 1-24.

Allen Ginsberg is a member of technical staff at AT&T Bell Laboratories. His research interests include information retrieval, multimedia indexing, and machine learning. He recieved a PhD in computer science in 1986 and a PhD in philosophy in 1983, both from Rutgers University. He can be reached at AT&T Bell Laboratories, Room 4G-614, Crawfords Corner Road, Holmdel, NJ 07733; e-mail abg@research.att.com

A unified approach to automatic indexing and information retrieval

- *Ginsberg, A.*
AT&T; Bell Lab., Holmdel, NJ, USA

This paper appears in: **IEEE Expert [see also IEEE Intelligent Systems]**
On page(s): 46 - 56
Oct. 1993
Volume: 8 Issue: 5
ISSN: 0885-9000
References Cited: 10
CODEN: IEEXE7
INSPEC Accession Number: 4562664

Abstract:
WorldViews, an experimental system that unites as many aspects of information organization as possible around a simple, familiar, yet versatile knowledge representation framework, is discussed. This framework is a lattice-structured version of the traditional thesaurus. WorldViews focuses on the unified use of the thesaurus to automatically index and retrieve information. The performance of the WorldViews system is outlined.

Index Terms:
automatic indexing; information retrieval; experimental system; information organization; knowledge representation framework; lattice-structured version; WorldViews; unified use; indexing; information retrieval systems; knowledge based systems; knowledge representation; thesauri